A bootstrap strategy for the detection of a panel attrition bias in a household panel with an application to the German Socio-Economic Panel (GSOEP)

Rendtel, Ulrich; Büchel, Felix

A Bootstrap Strategy for the Detection of a Panel Attrition Bias in a Household Panel with an Application to the German Socio–Economic Panel (GSOEP)

ULRICH RENDTEL AND FELIX BÜCHEL

Abstract: *A bootstrap strategy for detecting non–ignorable panel attrition is proposed. The strategy is based on the difference of the original estimate and an estimate that is obtained by reducing the original sample by a second attrition experiment. The attrition propensities are estimated from previous wave information and field work information of the current panel wave. The routine may be used to estimate the bias due to panel attrition. The bootstrap routine is applied to two income estimates with data from the first 8 waves of the GSOEP.*

Keywords: *attrition bias, bootstrap, household panel*

1 Introduction

Panel surveys are plagued by the successive attrition of people, who refuse to continue to participate or who are lost because of problems in recontacting them in the next wave of the panel. Such losses not only reduce the sample size, they may also bias estimates based on the remaining sample. The panel attrition is ignorable if conditioning on the participation does not affect the distribution of interest $f(Y \mid X)$; which means we have $f(Y \mid X, S = 1) = f(Y \mid X, S = 0)$, where $S = 1$ indicates participation and $S = 0$ indicates attrition. A selection rule is called non–ignorable, if the these two distributions differ. A recent survey on non–ignorable panel attrition was presented by Verbeek and Nijman (1996).

The main difficulty in the treatment of attrition is the lack of knowledge about $f(Y \mid X, S = 0)$. In fact, without any knowledge about $f(Y \mid X, S = 0)$ the problem is not solvable on the basis of the observed data alone. In a panel, however, there exists a lot of information about attriters. The information arises from the characteristics observed in the previous panel waves. Also the field work of the present panel wave produces relevant information; namely, whether the person or the household has moved since the last interview or whether the

household has split up into two separate households. In the case of the German Socio-Economic Panel (GSOEP), an ongoing household panel started in 1984 (cf. Wagner et al. 1993a + b), field-related characteristics turned out to be the most relevant indicators for explaining drop-out during the panel (cf. Rendtel 1990, 1995). For a panel study which is based on face to face interviews, this is a plausible finding.

Such information may be used to estimate drop-out probabilities for panel members. The attrition probabilities help synthesize our knowledge about the attrition process using information on attriters and non-attriters. In order to answer the question of whether panel attrition affects the estimation of the model of interest, it is crucial to exploit the relationship between the variables of the model of interest and the characteristics from field work.

The attrition bias is defined here as follows: let $\hat{\theta}$ be an estimator of some parameter θ that characterizes the distribution $f(Y \mid X, \theta)$. For each unit we observe covariates Z that predict attrition, which is indicated by S. Let $\hat{\theta} = \hat{\theta}(X, Y)$ be the estimate of θ on the basis of the sample in the absence of attrition. Of course, we cannot observe $\hat{\theta}$. Denote by $\tilde{\theta} = \hat{\theta}(\tilde{X}, \tilde{Y})$ the estimate of θ on the basis of the observed sample $\tilde{X}, \tilde{Y}$ after attrition. We use here the following definition of an attrition bias of $\hat{\theta}$:

$$\text{bias}(\hat{\theta}) = E_S(\hat{\theta}(X, Y) - \hat{\theta}(\tilde{X}, \tilde{Y}) \mid X, Y, Z)$$

Hence, the bias($\hat{\theta}$) is the expected difference of the estimation results with and without attrition. The expectation is with respect to S conditional on the value of X , Y and Z. The effect of an attrition rule depends strongly on the marginal distribution of the model variables and the attrition predictors in the sample before attrition, which is reflected by conditioning on X,Y and Z .

The basic idea of the bootstrap strategy presented here is to resample from the observed sample $\tilde{X}, \tilde{Y}$ B replicates X^*, Y^*. This is done by Poisson sampling, which means that each unit is resampled according to its propensity of non-attrition. The use of the Poisson sampling is different from the standard Bootstrap routines which use sampling with replacement(Efron and Tibshirani 1993). The Poisson sampling results in a second, artificial attrition experiment on those units that survived the first attrition.

The bootstrap strategy presented here is also different from Efron's (1994) "Full-mechanism Bootstrap", since we do not try to reconstruct the distribution of X and Y before attrition.

Under some regularity conditions we may assume that the second attrition experiment produces a similar bias as the first attrition experiment. In this case we can check whether the distribution of the $\hat{\theta}^* = \hat{\theta}(X^*, Y^*)$ is centered around

$\widetilde{\theta} = \hat{\theta}(\tilde{X}, \tilde{Y})$. The average of the differences $\widetilde{\theta} - \hat{\theta}^*$ is taken as an estimate of bias$(\hat{\theta})$.

The bootstrap routine is applied to data from the first 8 waves of the GSOEP. The example deals with males who experienced a period of unemployment of at least 12 months. We want to know whether a joint analysis of incomes before and after unemployment is affected by panel attrition.

2 Selection on observable and unobservable variables

A necessary condition for the bootstrap strategy to work is the selection on observable variables. The distinction between selection on observable and selection on unobservable variables is discussed here within the the framework of the standard econometric model for selection. Here, we have a regression equation:

$$Y_{i,t} = X'_{i,t}\beta + \epsilon_{i,t} \tag{1}$$

which is observed if $S_{i,t} = 1$. Here i indicates units (individuals or households) and t indicates the panel wave. The model for $S_{i,t}$ is the stochastic censoring model for a latent response propensity $S^*_{i,t}$:

$$S^*_{i,t} = Z'_{i,t}\gamma + \delta_{i,t} \tag{2}$$

and

$$S_{i,t} = \begin{cases} 1 & \text{if } S^*_{i,t} \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

The standard orthogonality assumptions are: $X_{i,t} \perp \epsilon_{i,t}$, $Z_{i,t} \perp \delta_{i,t}$. Selection on unobservables holds, if:

$$E(\epsilon_{i,t} Z_{i,t} \mid X_{i,t}) = 0 \text{ and } E(\epsilon_{i,t}\delta_{i,t} \mid X_{i,t}) \neq 0 \tag{4}$$

This case applies if all Z–variables that are not regression covariates have no impact on $Y_{i,t}$. However, there exist unobserved variables that affect both the regression equation and the selection equation.

In the observable selection case the above relationship is interchanged. Here we have:

$$E(\epsilon_{i,t} Z_{i,t} \mid X_{i,t}) \neq 0 \text{ and } E(\epsilon_{i,t}\delta_{i,t} \mid X_{i,t}) = 0 \tag{5}$$

This case applies if some of the Z–variables that are not regression variables have an impact on $Y_{i,t}$ which is not explained by the covariates of the regression model.

However, the model ignores the existence of unobserved variables that affect both, the regression equation and the selection equation. The observable selection appears to be well suited for the case in which $Z_{i,t}$ contains lagged dependent variables like $Y_{i,t-1}$ or $Y_{i,t-2}$. The occurence of $Y_{i,t-1}$ as a covariate for the attrition propensity scores is quite natural in a panel since $Y_{i,t-1}$ is in general known for all attriters in wave t. However, it is unrealistic to assume $E(\epsilon_{i,t}Z_{i,t} \mid X_{i,t}) = 0$ as long as $Y_{i,t-1}$ is not a covariate in the regression model.

It is immediately clear that the bootstrap strategy will fail to detect an attrition bias in case of the selection on unobservables: the resampling strategy explicitely uses the independence of $\epsilon_{i,t}$ and $\delta^*_{i,t}$, which is the simulated random propensity part in equation 2.

However, also the selection on unobserved variables makes strong asumptions about the joint distribution of $X_{i,t}, Y_{i,t}$ and $Z_{i,t}$. It assumes that the regression coefficient of Z–variables not contained in $X_{i,t}$ are 0 in the unselected population. Such a restriction cannot be tested on the basis of the observed data. It has to be deduced from a–priori knowledge. For example, if the survey organisation introduces some random variation of fieldwork rules which are known to have different impacts on the participation behavior, the fieldwork treatment indicator can be guaranteed to have no impact on Y but surely it has an influence on the participation behavior. This happened, for example, in the British household panel survey (BHPS), where a change of the interviewer was randomly introduced[1]. Such changes are known to be a source of an increased attrition risk (Rendtel 1990,1995).

However, such experimental rules of fieldwork are expensive (also with respect to attrition rates) and therefore seldom. Fitzgerald et al. (1997) conclude in their analysis of sample attrition in the PSID "that there are no suitable candidates for instruments for nonresponse[2] in the PSID and hence that we cannot adjust for selection on unobservables".

3 The implementation of the bootstrap procedure

The implementation of the bootstrap procedure is much facilitated if the data base contains variables with the estimated propensity to attrit in the current wave given participation in the proceeding wave [3]. In order to replicate the original attrition

[1]Such a rule is different from changes of the interviewer that are caused be the move of a household.

[2]i.e. Z–variables where a correlation with Y can be excluded by a-priori knowledge.

[3]The GSOEP data base containes variables which describe the reciprocal value of the risk that a household attrits from the preceeding wave to the current wave. Details that decribe their generation can be found in Rendtel (1995)

process the following rules appear to be appropriate:

1. Since the attrition risk occurs sequentially from one wave to another, the bootstrap attrition should also be performed sequentially. This is especially useful if the estimation procedure bases on an unbalanced sample.
2. In many panels attrition is an absorbing state, i.e. attrited units do not re–enter the panel. Consequently the bootstrap attrition should be absorbing. Therefore,after an attrition has occured all observations of the unit and their household splitt–offs after that wave should be skipped in the bootstrap routine.
3. One should also reflect dependencies in the participation of household members. As a rule, household members react unanimously, i.e. all household members cooperate or refuse their cooperation, see Rendtel (1995) for empirical results from the GSOEP. This strictly votes for an application of the bootstrap attrition at the household level.

4 An application of the bootstrap attrition routine

In this section we use the bootstrap approach for a sample from the German Socio–Economic Panel (GSOEP). The GSOEP is an ongoing household panel, which is similar to the Panel Study of Income Dynamics (PSID). It started in 1984 with a sample of about 6000 households and 12000 interviewed persons. A short description of the data base is given in Wagner et al. (1993a). Detailed information can be found in Wagner et al. (1993b) and Haisken-DeNew and Frick (1997).

All household members who are older than 16 years are interviewed. The regular interviewing method is a personal interview or a self–filled questionaire in the presence of the interviewer. All persons that have given an interview are followed up as long as they stay within Germany.

The GSOEP was at its 13^{th} wave in 1996. Up to that time it had lost more than 40% of its wave 1 members through panel attrition. Because wages and labor force participation are the central topics of the GSOEP they have been choosen to be checked for effects of panel attrition.

The model used in this section is the basic model of human capital theory, which explains the log of the earned monthly gross income by the duration of the education (schooling), the duration of the participation in the labor force (experience) and the firm specific human capital expressed by the length of the job at the present employer (tenure). Such a model was also used by Becketti et al. (1988) and Fitzgerald et al. (1996) to evaluate the PSID[4].

[4]Becketti et al. did not use tenure. Instead they used some race dummies.

The choice of our example was motivated by the following considerations:

- The selected group should have a high potential risk of attrition.
- The aim of the analysis should be panel specific, for example, a before/after event comparison.

We have chosen here the group of male employees that experienced a period of unemployment of at least 12 months. The risk of attrition of these people is supposed to be high because they are less educated and have a higher propensity of residential mobility in order to get a new job. The aim of the analysis is to assess the effects of unemployment on earned income[5] if we control for basic variables such as schooling, experience and tenure.

In this example the human capital model is augmented by interactions with the indicator LTU for observations after the long–term unemployment period. Hence, the coefficient of LTU*schooling measures the depreciation of school–specific human capital, while LTU*experience and LTU*(experience^2) describe the depreciation of occupationally achieved human capital. We did not include an interaction term with tenure since the firm–specific human capital is usually completely devalued after long–term unemployment[6].

In order to reduce the number of parameters we deflated the gross income. Therefore we used only one time dummy measuring real wage increases after 1987, the year that marks the end of the economic recession in the reference period in Germany.

The sample consists of 224 male employees with 761 valid income observations. Figure 1 shows the sample status of these people during the first eight waves of the GSOEP. It appears that there are to be almost no losses due to panel attrition during the first 3 waves of the panel, contradicting general knowledge that the panel attrition is highest at the beginning of the panel. The reason for this discrepancy arises from the fact that in most cases it is necessary to observe a person in two subsequent waves to assess a long–term unemployment period. Those people who are unemployed for less than a year and attrit are not in the sample.

[5]Therefore we excluded people from the analysis, where no before or after income exists, i.e. people who enter the labor force or retire.

[6]We assume that employees do not return to their old firms, which is the usual pattern of re–entering into the labor force after unemplovment in Germany.

Figure 1: Participation behavior during the panel: Male employees with long–term unemployment. Waves 1 to 8 of the GSOEP.

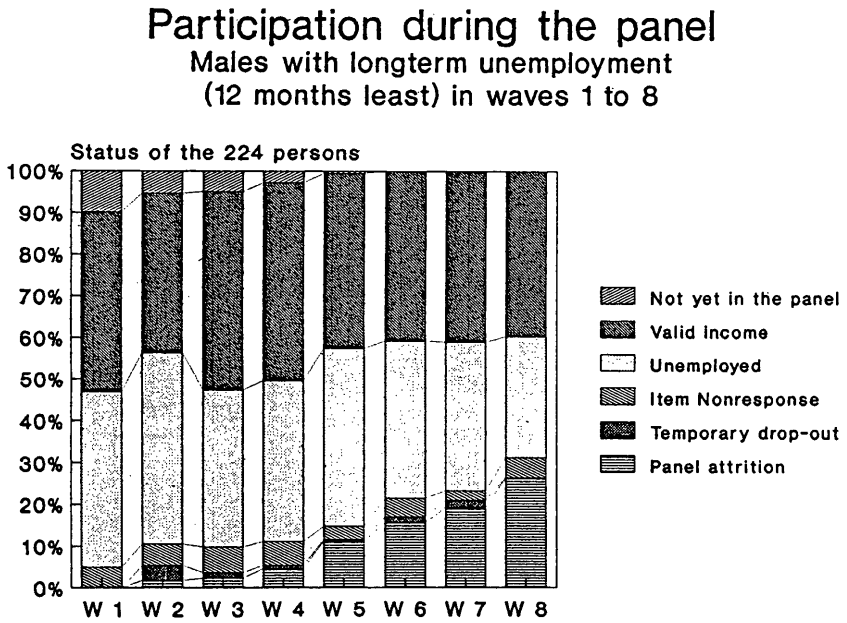


Table 1 displays the results of an unbalanced Feasable Generalized Least Squares (FGLS) analysis of a random effects model, which assumes an individual specific variance component for the error terms (Hsiao 1986, p.34). In this model all interactions of LTU with the other covariates turn out to be insignificant. Thus, there appears no further depreciation of human capital after long–term unemployment. However, there remain permanent effects due to at least one year of missing experience.

Table 1: Estimation of the income of male employees with an observed period of long–term unemployment (LTU) of at least 12 months. Dependent variable: ln(monthly gross income). Source: Waves 1 to 8 of the GSOEP. Number of persons: 224. Number of valid income measurements: 761.

Characteristic	FGLS estimate	t-values
Constant	7.261	50.9
After 1987	0.047	2.2
Schooling	0.031	2.5
Experience	0.020	2.4
Experience^2	-3.6×10^{-4}	-1.6
Tenure	0.006	2.1
After LTU	0.132	0.9
After LTU*Schooling	-0 007	-0.6
After LTU*Experience	-0.004	-0.5
After LTU*(Experience^2)	0.3×10^{-4}	0.1

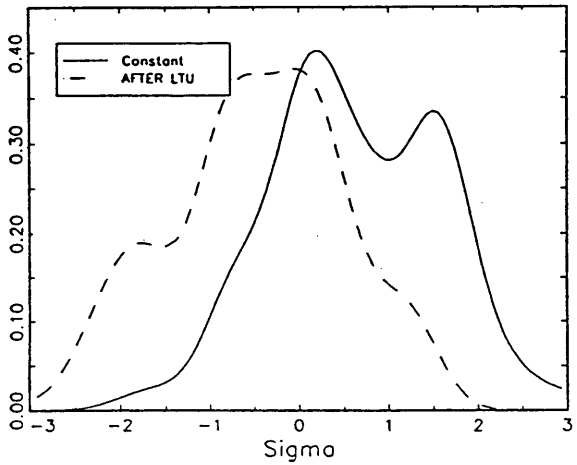
In Table 2 we compare the means and standard deviations of $\widetilde{\beta} - \hat{\beta}^*$. We find no indication of a large systematic bias. The highest bias is indicated for the constant and for the interaction of LTU with the constant. The distribution of $\widetilde{\beta} - \hat{\beta}^*$ for these two coefficients is displayed in Figure 2. There appears a bimodality for the simulated distribution of both coefficients and it is obvious that the minor modal values correspond to each other[7]. Such a plausible correspondence indicates that the estimated trade–off of both coefficients is sensitive to the panel attrition.

[7]For the distribution of the other coefficients, not documented here, there is no such bimodality.

Table 2: Results of B=100 bootstrap attrition experiments. $\widetilde{\beta}$: Estimate of β on the basis of the observed sample. $\hat{\beta}^*$: Estimate of β on the basis of the bootstrap sample.

Characteristic	Mean of $\widetilde{\beta} - \hat{\beta}^*$	Std. Deviation of $\widetilde{\beta} - \hat{\beta}^*$
Constant	0.0437	0.0650
After 1987	-0.0090	0.0126
Schooling	-0.0025	0.0057
Experience	-0.0021	0.0049
Experience^2	0.4×10^{-4}	1.1×10^{-4}
Tenure	-0.0004	0.0018
After LTU	-0.0577	0.1173
After LTU*Schooling	0.0051	0.0113
After LTU*Experience	0.0019	0.0058
After LTU*(Experience^2)	-0.5×10^{-4}	1.4×10^{-4}

Figure 2: The distribution of $\widetilde{\beta} - \hat{\beta}^*$ for the constant and the coefficient LTU. Kernel density estimate on the basis of B=100 bootstrap attrition experiments.



5 Conclusions

In many panel studies there are good reasons to assume that panel attrition is caused by field–related variables. If the model variables are not correlated with the field work variables attrition will turn out to be ignorable.
The bootstrap routine we proposed here efficiently uses the available knowledge about the attrition process in a panel study. The attractive features of the routine are:

- It is not necessary for the reseacher to estimate an attrition model if there are appropriate variables with attrition propensities in the data base.
- The researcher does not have to use a new estimation routine for his model of interest. It suffices to apply the same estimation routine to different data sets.
- The routine works for every analysis, not only for regression analysis.
- The routine gives a reasonable estimate of the size of an attrition bias.

All the researcher has to do is the programming and the execution of the bootstrap attrition experiments, which seems relatively simple.

In order to achieve these merits one has to rely on the assumption that the selection process can be controlled by observable variables and that these observable variables are contained in the model underlying the generation of the attrition propensities in the data base.

There are some alternatives to the bootstrap routine, especially in the case of regression analysis: First, one can augment the regression equation with the observed attrition variables. However, nonzero estimated slope coefficients of these variables are not always an indicator of an attrition bias nor differences in the estimated β–values indicate always an attrition bias.

Second, one can run a weighted regression analysis where the weights are generated by the inverse of the attrition propensities. Under the assumption of selection on observables such a weighted regression analysis yields consistent estimates of β, see Cosslett (1993), DuMouchel and Duncan (1983), Fitzgerald et al. (1996), Little (1991) and Nathan and Holt (1980).

References

Becketti, Sean; Gould, William; Lillard, Lee; Welch, Finis (1988). The Panel Study of Income Dynamics After 14 Years: An Evaluation. Journal of Labor Economics, 6, 472–492.

Cosslet, S. (1993). Estimation from Endogenously Stratified Samples. In: Handbook of Statistics, Vol.11, eds. Maddala, Rao, Vinod, Elsevier.

DuMouchel, William; Duncan, Greg (1983). Using Sample Survey Weights in Multiple Regression Analysis of Stratified Samples. Journal of the American Statistical Association, 78, 535–543.

Efron, Bradley (1994). Missing Data,Imputation, and the Bootstrap, Journal of the American Statistical Association, 89, 463–475.

Efron, Bradley; Tibshirani, R. (1993). An Introduction to the Bootstrap, Chapman and Hall,London.

Fitzgerald, John; Gottschalk, Peter; Moffitt, Robert (1996). An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics, Paper presented at the 7^{th} Conference on Panel Data, Paris.

Haisken–Denew, John; Frick, Joachim (1997). Destop Companion to the German Socio–Economic Panel Study (GSOEP), German Institut for Economic Research (DIW), Berlin.

Hsiao, Cheng (1986). Analysis of Panel Data, Cambridge University Press, Cambridge.

Little, Roderick (1991). Inference with Survey Weights. Journal of Official Statistics, 7, 405–424.

Nathan, G. ;Holt, D. (1980). The Effect of Survey Design on Regression Analysis. Journal of the Royal Statistical Society, B, 42, 377–386.

Rendtel, Ulrich (1990). Teilnahmeentscheidung in Panelstudien: Zwischen Beeinflussung, Vertrauen und sozialer Selektion. Über die Entwicklung der Antwortbereitschaft im Sozio–ökonomischen Panel. Kölner Zeitschrift für Soziologie und Sozialpsychologie, 42, 280–299.

Rendtel, Ulrich (1995). Panelausfälle und Panelrepräsentativität, Campus Verlag, Frankfurt/M. New–York.

Verbeek, Marno; Nijman, Theo (1996) Incomplete Panels and Selection Bias: A Survey. In: Matyas, L.; Sevestre, P. (eds): The Econometrics of Panel Data: Theory and Applications, $2^{n}dEdition$, Kluwer Academic Publishers, Dordrecht London.

Wagner, Gert; Burkhauser, Richard; Behringer, Friederike (1993a). The English Language Public Use File of the German Socio–Economic Panel. Journal of Human Resources, 28, 429–433.

Wagner, Gert; Schupp, Jürgen; Rendtel, Ulrich (1993b). Das Sozio–oekonomische Panel (SOEP) — Methoden der Datenproduktion und –aufbereitung im Längsschnitt. In: Hauser, R.; Ott, N.; Wagner, G. (Hrsg.): Mikroanalytische Grundlagen der Gesellschaftspolitik — Band 2: Erhebungsverfahren, Analysemethoden und Mikrosimulation, Akademie Verlag, Berlin, 70–112.